CHAPTER 2

STUDY AND REVIEW OF LITERATURE

2.1 Introduction

2.2 Concepts and Terminology of Data Mining

   2.2.1 Exploring Key Terms of Data Mining

      2.2.1.1 Knowledge Discovery in Database

      2.2.1.2 Data Warehouse

      2.2.1.3 Data Cube

   2.2.2 Primitives of Data Mining

   2.2.3 Classification of Data Mining Techniques

2.3 Applications of Data Mining

2.4 Review of Existing use of Data Mining Techniques

2.5 Summary

2.1 INTRODUCTION

Prevention and Detection of financial statement fraud should complement and supplement each other. However, [Bologna95] states that prevention should take precedence over detection. Auditing procedures are not skilled for prevention and detection of financial statement fraud, because it is not their key objective. Ethically, management of an organisation is supposed to be accountable for prevention and detection of financial statement fraud but it is almost always accomplished with the consent or knowledge of management. Failure to detect or prevent financial statement fraud can damage the reputation and the credibility of the audit profession [Chui11]. In order to help auditors, analytical techniques of data mining can estimate the risk of fraud and can help them in understanding the reasons behind the fraudulent financial reporting.
Hence, various techniques of data mining are being used to ease out this extra pressure of prevention and detection of financial statement fraud, from the mind of the auditors.

Data mining is a confluence of the various disciplines such as statistics, artificial intelligence, and pattern recognition. With the coming of data mining as a new field of data analysis, data analysing techniques can be divided into two groups: reporting techniques and data mining techniques. Reporting techniques refers to the techniques used before, where quantitative and statistical data characteristics are extracted from data and human analysts turn this information into knowledge. These are the techniques currently used in internal control settings. Data mining techniques emphasizes on the semi-automatic process to discover meaningful patterns in large data sets. Especially the data mining characteristic of revealing latent knowledge is very typical and valuable. This characteristic comes forward in the fact that no hypotheses are needed to mine the data, as opposed to pure statistics or data reporting [Jans09]. This is the main reason why these techniques are selected for preventing and detecting financial statement fraud in this research.

Data mining techniques can be used for assisting auditors in prevention and detection of financial statement fraud because these methods are capable of constructing self learned models from historical cases of fraud, which identify and detect the risk of fraud. Data Mining is an iterative process within which progress is defined by discovery of knowledge, which helps in finding the reasons behind financial statement fraud. Data Mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an “interesting” outcome [Kantardzi02].

The application of Data Mining techniques for detection and identification of financial statement fraud is a fertile research area. Several law enforcement agencies and special investigative units have used data mining techniques successfully for detection of financial frauds.

Traditional methods of auditing and internal control are not capable enough for prevention and detection of financial statement fraud, because it is a type of management fraud and management is adaptive and usually find easy ways to circumvent the auditing measures. Therefore, several data mining techniques have been implemented by number of researchers for preventing and detecting fraudulent financial reporting. In this research work extensive literature survey is carried out in the area of applicability of data mining methods for prevention and detection of financial statement fraud by focusing on nature of data mining techniques and data specifications.
In addition various empirical results of data mining techniques are also investigated. The literature survey is presented in this chapter under Section 2.4. Before presenting the survey elementary concepts and terminology of data mining is given in section 2.2 which presents the process of knowledge discovery, fundamental concepts of data mining – data warehouse and data cube, primitives and classification of data mining methods. Section 2.3 presents the various applications of data mining. The chapter is summarised in Section 2.5.

2.2 CONCEPTS AND TERMINOLOGY OF DATA MINING

2.2.1 Exploring Key Terms of Data Mining

Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. The process of knowledge discovery is explained below:

2.2.1.1 Knowledge Discovery in Database

The current information age is overwhelmed by data. More and more information is stored in databases and turning these data into knowledge creates a demand for new, powerful tools. Data analysis techniques used before were primarily oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. These interpretations and insights are the sought knowledge. Although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analysts [Michalski98]. The current situation however needed a new way to deal with these never ending databases and new methods to analyse this huge amount of data. A new area came into being: Knowledge Discovery in Databases, also known as KDD. The process of KDD is depicted as Figure 2.1 and consists of an iterative sequence of the following steps [Han01]:

1. Data cleaning

Real-world data tend to be incomplete, noisy, and inconsistent. Dirty data can cause confusion for the mining procedure. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding over fitting the data to the function being modelled. Therefore, a Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.
2. **Data integration**

The task of data integration combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files.

3. **Data selection**

The task of data selection includes retrieving data relevant to the analysis task from the database.

4. **Data transformation**

Data are transformed or consolidated into forms appropriate for mining by performing the following:

   a. **Smoothing**: It works to remove the noise from data. Such techniques include binning, clustering, and regression.

   b. **Aggregation**: It means applying summary or aggregation operations to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.

   c. **Generalization** of the data, where low level or primitive (raw) data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher level concepts, like city or county. Similarly, values for numeric attributes, like age, may be mapped to higher level concepts, like young, middle-aged, and senior.

   d. **Normalization**, where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.

   e. **Attribute construction (or feature construction)**, where new attributes are constructed and added from the given set of attributes to help the mining process.

5. **Data mining**

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both. Data mining is the process of finding
correlations or patterns among dozens of fields in large relational databases. Data mining is a powerful tool because it can provide you with relevant information that you can use to your own advantage.

Data mining is a logical process that is used to search through large amounts of information in order to find important data. The goal of this technique is to find patterns that were previously unknown. It is an essential process where intelligent methods are applied in order to extract data patterns

6. Pattern evaluation

This step is required to identify the truly interesting patterns representing knowledge based on some interestingness measures. Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.
- Task-relevant Data
- Data Warehouse
- Data Cleaning
- Data Integration
- Databases
- Pattern Evaluation

Figure 2.1: The Process of Knowledge Discovery
7. Knowledge presentation

This final step includes the use of visualization and knowledge representation techniques to present the mined knowledge to the user. The discovered knowledge must be organized and presented in a way that the user can use. Depending on the requirements, this step can be as simple as generating a pie or a bar chart or as complex as generating a Decision Tree.

It is clear from this figure that an integral part of the process of KDD is data mining. Together with KDD, data mining was born as a new research field. Data mining is a reaction to overcome the above limitations of data analyzing techniques used before. A data analysis system now has to be equipped with a substantial amount of background knowledge, and be able to perform reasoning tasks involving that knowledge and the data provided [Han01].

The first two steps of KDD process namely Data Cleaning and Data Integration can be viewed as an important preprocessing step for data mining. These two steps results in a data warehouse which is explained as follows:

2.2.1.2 Data Warehouse

The construction of data warehouses, which involves data cleaning and data integration, can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Furthermore, many other data mining functions such as classification, prediction, association, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction. Hence, data warehouse has become an increasingly important platform for data analysis and online analytical processing and will provide an effective platform for data mining [Han01].

A data warehouse is a subject-oriented, integrated, time-variant, and non – volatile collection of data in support of management's decision making process [Inmon95]. These four keywords distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems. The four key features of data warehouse are explained below:
a. **Subject-oriented:** A data warehouse is organized around major subjects, such as customer, vendor, product, and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modelling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

b. **Integrated:** A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

c. **Time-variant:** Data are stored to provide information from a historical perspective (e.g., the past 5-10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

d. **Non-volatile:** A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

In sum, a data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions.

Data warehouses are based on a multidimensional data model. A multidimensional view of data is in the form of a data cube which is described below:

### 2.2.1.3 Data Cube

A data cube allows data to be modelled and viewed in multiple dimensions. It is defined by dimensions and facts [Han01].

In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records. For example, an organization may create a sales data warehouse in order to keep records of their sales with respect to the dimensions time, item, branch, and location.
These dimensions allow the store to keep track of things like monthly sales of items, and the branches and locations at which the items were sold. Each dimension may have a table associated with it, called a dimension table, which further describes the dimension. For example, a dimension table for item may contain the attributes item name, brand, and type. Dimension tables can be specified by users or experts, or automatically generated and adjusted based on data distributions.

A multidimensional data model is typically organized around a central theme, like sales, for instance. This theme is represented by a fact table. Facts are numerical measures. Think of them as the quantities by which we want to analyse relationships between dimensions. Examples of facts for a sales data warehouse include Rupees sold (sales amount in rupees), units sold (number of units sold), and amount budgeted. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

2.2.2 Primitives of Data Mining

The users of data mining systems can communicate with it by using a set of data mining primitives designed in order to facilitate efficient and fruitful knowledge discovery. Such primitives include the specification of the portions of the database or the set of data in which the user is interested, the kinds of knowledge to be mined, background knowledge useful in guiding the discovery process, interestingness measures for pattern evaluation, and how the discovered knowledge should be visualized. These primitives allow the user to interactively communicate with the data mining system during discovery in order to examine the findings from different angles or depths, and direct the mining process. Brief overview of these data mining primitives is given below [Han01].

1. Task-relevant data

The first primitive is the specification of the data on which mining is to be performed. Typically, a user is interested in only a subset of the database. It is impractical to indiscriminately mine the entire database, particularly since the number of patterns generated could be exponential with respect to the database size. Furthermore, many of these patterns found would be irrelevant to the interests of the user.

2. The kinds of knowledge to be mined

This species the data mining functions to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. In addition to specifying the kind of knowledge to be mined for a given data mining task, the user can be more specific and provide pattern templates that all discovered patterns must match.
These templates, or meta – patterns also called meta – rules or meta – queries, can be used to guide the discovery process.

3. **Background knowledge**

Users can specify background knowledge, or knowledge about the domain to be mined. This knowledge is useful for guiding the knowledge discovery process, and for evaluating the patterns found. There are several kinds of background knowledge. One of the most popular forms of background knowledge is concept hierarchies. Concept hierarchies are useful in that they allow data to be mined at multiple levels of abstraction. Other examples include user beliefs regarding relationships in the data.

These can be used to evaluate the discovered patterns according to their degree of unexpectedness, where unexpected patterns are deemed interesting.

4. **Interestingness measures**

These functions are used to separate uninteresting patterns from knowledge. They may be used to guide the mining process, or after discovery, to evaluate the discovered patterns. Although specification of the task-relevant data and of the kind of knowledge to be mined may substantially reduce the number of patterns generated, a data mining process may still generate a large number of patterns. Typically, only a small fraction of these patterns will actually be of interest to the given user. Thus, users need to further confine the number of uninteresting patterns returned by the process. This can be achieved by specifying interestingness measures which estimate the simplicity, certainty, utility, and novelty of patterns.

Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support (the percentage of task-relevant data tuples for which the rule pattern appears), and confidence (the strength of the implication of the rule). Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

5. **Presentation and visualization of discovered patterns**

This refers to the form in which discovered patterns are to be displayed. Users can choose from different forms for knowledge presentation, such as rules, tables, charts, graphs, decision trees, and cubes.

For data mining to be effective, data mining systems should be able to display the discovered patterns in multiple forms, such as rules, tables, crosstabs, pie or bar charts, decision trees, cubes, or other visual representations.
Allowing the visualization of discovered patterns in various forms can help users with different backgrounds to identify patterns of interest and to interact or guide the system in further discovery.

A user should be able to specify the kinds of presentation to be used for displaying the discovered patterns. Some representation forms may be better suited than others for particular kinds of knowledge. For example, generalized relations and their corresponding crosstabs (cross-tabulations) or pie/bar charts are good for presenting characteristic descriptions, whereas decision trees are a common choice for classification.

2.2.3 Classification of Data Mining Techniques

Data mining techniques can be divided into two subgroups: Descriptive and predictive data mining techniques. Predictive data mining analyses the data in order to construct one or a set of models, and attempts to predict the behaviour of new data set whereas descriptive data mining describes the dataset in a concise and summative manner and presents interesting general properties of data. [Han01].

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high performance computing.

Because of the diversity of disciplines contributing to data mining, data mining research is expected to generate a large variety of data mining systems. Therefore, it is necessary to provide a clear classification of data mining systems. Such a classification may help potential users distinguish data mining systems and identify those that best match their needs. Data mining systems can be categorized according to various criteria, as follows [Han01].

2.2.3.1 Classification according to the kinds of databases mined.

A data mining system can be classified according to the kinds of databases mined. Database systems themselves can be classified according to different criteria such as data models, or the types of data or applications involved, each of which may require its own data mining technique. Data mining systems can therefore be classified accordingly. For instance, if classifying according to data models, we may have a relational, transactional, object-oriented, object-relational, or data warehouse mining system.
If classifying according to the special types of data handled, we may have a spatial, time-series, text, or multimedia data mining system, or a World-Wide Web mining system. Other system types include heterogeneous data mining systems, and legacy data mining systems.

2.2.3.2 **Classification according to the kinds of knowledge mined.**

Data mining systems can be categorized according to the kinds of knowledge they mine, i.e., based on data mining functionalities, such as characterization, discrimination, association, classification, clustering, trend and evolution analysis, deviation analysis, similarity analysis, etc. A comprehensive data mining system usually provides multiple and/or integrated data mining functionalities.

Moreover, data mining systems can also be distinguished based on the granularity or levels of abstraction of the knowledge mined, including generalized knowledge (at a high level of abstraction), primitive-level knowledge (at a raw data level), or knowledge at multiple levels (considering several levels of abstraction). An advanced data mining system should facilitate the discovery of knowledge at multiple levels of abstraction.

Based on the type of knowledge that is mined, data mining can be mainly classified into the following categories [Zhang04].

1) **Association rule mining** uncovers interesting correlation patterns among a large set of data items by showing attribute-value conditions that occur together frequently. A typical example is market basket analysis, which analyzes purchasing habits of customers by finding associations between different items in customers' "shopping baskets."

2) **Classification and prediction** is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. The models are used to predict the class of objects whose class label is unknown. A bank, for example, may classify a loan application as either a fraud or a potential business using models based on characteristics of the applicant. A large number of classification models have been developed for predicting future trends of stock market indices and foreign exchange rates.

3) **Clustering analysis** segments a large set of data into subsets or clusters. Each cluster is a collection of data objects that are similar to one another within the same cluster but dissimilar to objects in other clusters. In other words, objects are clustered based on the principle of maximizing the intra-class similarity while minimizing the inter-class similarity. For
example, clustering techniques can be used to identify stable dependencies for risk management and investment management.

4) *Sequential pattern and time-series mining* looks for patterns where one event (or value) leads to another later event (or value). One example is that after the inflation rate increases, the stock market is likely to go down.

The knowledge to be mined is closely related to a target application and the original data. Therefore, data mining should be considered along with several other issues rather than an isolated task.

### 2.2.3.3 Classification according to the kinds of techniques utilized.

Data mining systems can also be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved e.g., autonomous systems, interactive exploratory systems, query-driven systems, or the methods of data analysis employed e.g., database-oriented or data warehouse-oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks, and so on. A sophisticated data mining system will often adopt multiple data mining techniques or work out an effective, integrated technique which combines the merits of a few individual approaches.

### 2.3 APPLICATIONS OF DATA MINING

Data Mining is a process that analyzes the large amount of data to find the new and hidden information that improves business efficiency. Various industries have been adopting data mining to their mission-critical business processes to gain competitive advantages and help business grows. This section illustrates some data mining applications in Finance, Retail Industry, and telecommunication industry.

#### 2.3.1 Data Mining Applications in Finance

##### 2.3.1.1 Detection of money laundering and other financial crimes

To detect money laundering and other financial crimes, it is important to integrate information from multiple databases (like bank transaction databases, and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers.
Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of customers), linkage analysis tools (to identify links among different customers and activities), classification tools (to filter unrelated attributes and rank the highly related ones), clustering tools (to group different cases), outlier analysis tools (to detect unusual amounts of fund transfers or other activities), and sequential pattern analysis tools (to characterize unusual access sequences). These tools may identify important relationships and patterns of activities and help investigators focus on suspicious cases for further detailed examination.

2.3.1.2 Classification and clustering of customers for targeted marketing

Classification and clustering methods can be used for customer group identification and targeted marketing. For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking. Customers with similar behaviours regarding loan payments may be identified by multidimensional clustering techniques. These can help identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.

2.3.1.3 Loan payment prediction and customer credit policy analysis

Loan payment prediction and customer credit analysis are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus the total monthly income), payment-to-income ratio, customer income level, education level, residence region, and credit history. Analysis of the customer payment history may find that, payment-to-income ratio is a dominant factor, while education level and debt ratio are not. The bank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profile shows relatively low risks according to the critical factor analysis.

2.3.2 Data Mining Application in Retail Industry

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on the internet, or e-commerce. Retail data mining can help identify customer buying behaviours, discover customer shopping
patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business. A few examples of data mining in the retail industry are outlined as follows.

2.3.2.1 Analysis of the effectiveness of sales campaigns

The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers. Careful analysis of the effectiveness of sales campaigns can help in improving company’s profit. Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of transactions containing the sales items during the sales period versus those containing the same items before or after the sales campaign. Moreover, association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

2.3.2.2 Customer retention—analysis of customer loyalty

With customer loyalty card information, one can register sequences of purchases of particular customers. Customer loyalty and purchase trends can be analysed systematically. Goods purchased at different periods by the same customers can be grouped into sequences. Sequential pattern mining can then be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new ones.

2.3.2.3 Product recommendation and cross-referencing of items:

By mining associations from sales records, one may discover that a customer who buys a digital camera is likely to buy another set of items. Such information can be used to form product recommendations. Collaborative recommender systems use data mining techniques to make personalized product recommendations during live customer transactions, based on the opinions of other customers. Product recommendations can also be advertised on sales receipts, in weekly flyers, or on the Web to help improve customer service, aid customers in selecting items, and increase sales. Similarly, information such as “hot items this week” or attractive deals can be displayed together with the associative information in order to promote sales.

2.3.3 Data Mining for the Telecommunication Industry

The telecommunication industry has quickly evolved from offering local and long distance telephone services to providing many other comprehensive communication services,
including fax, pager, cellular phone, Internet messenger, images, e-mail, computer and Web data transmission, and other data traffic. The integration of telecommunication, computer network, Internet, and numerous other means of communication and computing is also underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service. The following are a few scenarios for which data mining may improve telecommunication services:

2.3.3.1 Fraudulent pattern analysis and the identification of unusual patterns

Fraudulent activity costs the telecommunication industry millions of dollars per year. It is important to (1) identify potentially fraudulent users and their typical usage patterns; (2) detect attempts to gain fraudulent entry to customer accounts; and (3) discover unusual patterns that may need special attention, such as busy hour frustrated call attempts, switch and route congestion patterns, and periodic calls from automatic dial-out equipment (like fax machines) that have been improperly programmed. Many of these patterns can be discovered by multidimensional analysis, cluster analysis, and outlier analysis.

2.3.3.2 Mobile telecommunication services

Mobile telecommunication, Web and information services, and mobile computing are becoming increasingly integrated and common in our work and life. One important feature of mobile telecommunication data is its association with spatiotemporal information. Spatiotemporal data mining may become essential for finding certain patterns. For example, unusually busy mobile phone traffic at certain locations may indicate something abnormal happening in these locations. Moreover, ease of use is crucial for enticing customers to adopt new mobile services. Data mining will likely play a major role in the design of adaptive solutions enabling users to obtain useful information with relatively few keystrokes.

2.4 REVIEW OF EXISTING USE OF DATA MINING TECHNIQUES

Fraud has an incurred cost too. It has been estimated that the typical organization loses 5% of its revenue to fraud each year. The median loss caused by occupational fraud cases was $140,000. More than one – fifth of these cases caused losses of at least $1 million. Perpetrators with higher levels of authority tend to cause much larger losses. The median loss
among frauds committed by owner/executives was $573,000, the median loss caused by managers was $180,000 and the median loss caused by employees was $60,000 [ACFE12].

Cost of financial statement fraud is very high both in terms of finance as well as the goodwill of the organisation and related country. In order to curb the chances of fraud and to detect the fraudulent financial reporting, number of researchers had used various techniques from the field of statics, artificial intelligence and data mining. Several groups of researchers have devoted a significant amount of effort in studying the use of data mining techniques in detection of financial statements fraud from different perspectives.

The first proposal of applying data mining techniques for detection of financial statement fraud found in literature is Green's experiment. Since Green's proposal, large number of data mining methods has been proposed in the literature. [Green97] evaluated the performance of artificial neural networks in dividing organizations in to fraud or non-fraud. Many other data mining techniques are also implemented by many researchers. Some of the major researches are given below along with purpose, nature of data mining method, data specifications and experimental results.

2.4.1 Green et al. [Green97]

**Purpose:** To evaluate the performance of three Artificial Neural Networks with input variables pre-processed in different ways: simple percentage change, plain sum-of-the-years'-digit weighted average, and incremental sum-of-the-years'-digit weighted average.

**Data Mining Techniques used:** Artificial Neural Network

**Nature of Data Mining Techniques Used:** Predictive

**Data Used:** In an experiment with 86 SEC fraud cases matched with 86 non-fraud cases the ANNs were compared to random guessing, defined as Type I and Type II error rates of 0.5, and summed error rate below 1.

**Results Obtained:** The results showed that the ANNs performed better than random guessing on the training sample. On the evaluation sample, however, the ANNs did not perform significantly better in terms of either Type I or Type II errors. The summed error rate comparison did show that the ANNs performed significantly better than random guessing, but this comparison used classification results from a combined sample of both the training and evaluation samples.
Conclusion: A Neural Network learns the pattern of input data for a fraud and non-fraud sample during model training. A classification model created from the learned behaviour pattern is then applied to a test sample. Three models, using different expectation methods to develop data input, act as an investigation rule to classify financial statement data. Neural Network technology allows the development of pattern aggregation by simultaneously evaluating isolated analytical procedure (AP) expectations. During the preliminary stage of an audit, a financial statement classified as fraudulent signals the auditor to increase substantive testing during fieldwork.

2.4.2 Fanning et al [Fanning98]

Purpose: To compared preliminary results of a neural network approach with those of stepwise logistic regression, linear discriminant analysis and quadratic discriminant analysis for detecting fraud in the first year of fraudulent filing charged by the SEC.

Data Mining Technique used: Artificial Neural Network

Nature of Data Mining Technique used: Predictive

Data used: They used a matched-pairs sample of 102 fraud companies identified in SEC AAERs and 102 non-fraud companies matched on industry, fiscal year end and company size.

Results Obtained: The results of the model developed in this research suggests that there is potential in detecting finding FFS through analysis of public documents. Further this study suggests that ANN’s offer superior ability to standard statistical methods in detecting FFS. AutoNet offers the advantage of quickly developing models for analysis. The neural net model included the following variables: percentage of outside directors on the board, having a non-Big Six auditor, growth, ratios of accounts receivable to sales, net plant property and equipment to assets, debt to equity, and trend variables for greater than 10% increase in accounts receivable and gross margin. It accurately classified 69% of the fraud companies, while misclassifying 20% of the non-fraud companies in their training data, and accurately identified 66% of the fraud companies, while misclassifying 41% of the non-fraud companies in their hold-out sample.

Conclusion: The study employed a self-organizing Artificial Neural Network (ANN) AutoNet in conjunction with standard statistical tools to investigate the usefulness of publicly available predictors. The study results in a model with a high probability of detecting fraudulent financial statements on one sample. The study reinforces the validity and
efficiency of AutoNet as a research tool and provides additional empirical evidence regarding
the merits of suggested red flags for fraudulent financial statements. Authors concluded by
saying that the neural network approach is effective in detecting fraud using publicly
available information.

2.4.3 Bell et al [Bell00]

Purpose: To develop and test a logistic regression model that estimates the likelihood of
fraudulent financial reporting for an audit client, conditioned on the presence or absence of
several fraud-risk factors.

Data Mining Technique used: Logistic Regression

Nature of Data Mining Technique used: Predictive

Data used: The data sample consists of 77 fraud engagements and 305 non fraud
engagements drawn from KPMG's audit practice.

Results Obtained: The model correctly classified 80% of the fraud cases while only
misclassifying 11% of the non-fraud cases. The significant risk factors included in the final
model were: weak internal control environment, rapid company growth, inadequate or
inconsistent relative profitability, management places undue emphasis on meeting earnings
projections, management lied to the auditors or was overly evasive, the ownership status
(public vs. private) of the entity, and an interaction term between a weak control environment
and an aggressive management attitude toward financial reporting. The logistic model was
significantly more accurate than practicing auditors in assessing risk for the 77 fraud
observations. There was not a significant difference between model assessments and those of
practicing auditors for the sample of non-fraud cases.

Conclusion: The findings suggest that a relatively simple decision aid performs quite well in
differentiating between fraud and non-fraud observations. Practitioners might consider using
this model, or one developed using a similar procedure, in fulfilling the SAS No. 82
requirement to “assess the risk of material misstatement of the financial statements due to
fraud.”
2.4.4 Spathis et al [Spathis02]

**Purpose:** To explore the effectiveness of an innovative classification methodology in detecting firms that issue falsified financial statements (FFS) and the identification of the factors associated to FFS.

**Data Mining Technique used:** Multi-criteria decision aid (MCDA) and UTADIS classification method (UTilite’s Additives DIScriminantes)

**Nature of Data Mining Technique used:** Predictive

**Data used:** A sample of 76 Greek firms (38 with FFS and 38 non-FFS) described over ten financial ratios is used for detecting factors associated with FFS. This study compares firms with discovered FFS that are publicly revealed to firms that do not have publicly revealed FFS. For the non-FFS firms of the sample, no published indication of FFS behaviour was uncovered in a search of databases and the relevant auditors’ reports.

**Results Obtained:** At each replication of the jackknife approach, different FFS detection models are developed using the UTADIS method. These models incorporate both the complete set of financial ratios as well as the reduced set of four ratios selected using factor analysis. The results indicate that the proposed MCDA classification approach is quite efficient in discriminating between FFS and non-FFS firms, thus supporting the conclusion that such an approach can assist auditors in their practice.

**Conclusion:** The results indicate that the proposed MCDA methodology outperforms traditional statistical techniques which are widely used for FFS detection purposes. Furthermore, the results indicate that the investigation of financial information can be helpful towards the identification of FFS and highlight the importance of financial ratios such as the total debt to total assets ratio, the inventories to sales ratio, the net profit to sales ratio and the sales to total assets ratio.

2.4.5 Koskivaara E. [Koskivaara03]

**Purpose:** To introduce the Artificial Neural Network technology and reviews the literature on auditing ANN applications.

**Data Mining Technique used:** Artificial Neural Network

**Nature of Data Mining Technique used:** Predictive

**Data used:** A number of articles have surveyed journal articles on ANNs applied to business situations. Twenty-one articles either focusing on or connected to the auditing environment
were found. All these articles fit into analytical review (AR) procedures. AR procedures are techniques used to improve the efficiency of audits. Basically, in an AR procedure one compares expected relationships among data items to actual observed relationships. Most existing AR procedures investigate ratios and trends of financial data.

Results Obtained: Artificial neural network (ANN) based information systems are proposed as one possible solution as a support tool for auditors.

Conclusion: As information technological changes occur at an increasing rate, auditors must keep pace with these emerging changes and their impact on their client’s information processing systems as well as on their own audit procedures. This paper reviewed the current state of the ANN-applications connected to auditing purpose. The review is comprehensive but by no means exhaustive, given the fast growing nature of the literature. The main findings are summarised as follows: The main application areas were material errors, management fraud, and support forgoing concern decision. ANNs have also been applied to internal control risk assessment, audit fee, and financial distress problems.

2.4.6 Koh et al [Koh04]

Purpose: To explore and compare the usefulness of neural networks, decision trees and logistic regression in predicting a firm's going concern status.

Data Mining Technique used: Neural Network, Decision Trees, and Logistic Regression

Nature of Data Mining Technique used: Predictive

Data used: The sample data comprise financial ratios for 165 going concerns and 165 matched non-going concerns.

Results Obtained: The classification results indicate the potential usefulness of data mining techniques in a going concern prediction context. Further, the decision tree going concern prediction model outperforms the logistic regression and neural network models. A decision tree has been constructed in this study in order to predict the hidden problems in financial statements by examining the following six variables: quick assets to current liabilities, market value of equity to total assets, total liabilities to total assets, interest payments to earnings before interest and tax, net income to total assets, and retained earnings to total assets.
Conclusion: Data mining techniques such as neural networks and decision trees are powerful for analysing complex non-linear and interaction relationships, and hence can supplement and complement traditional statistical methods in constructing going concern prediction models.

2.4.7 Kirkos et al [Korkos05]

Purpose: To explores the effectiveness of Data Mining (DM) classification techniques in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS.

Data Mining Technique used: Decision Tree, Neural Network & Bayesian Belief Network

Nature of Data Mining Technique used: Predictive

Data used: Data Set contained data from 76 Greek manufacturing firms (no financial companies were included). Auditors checked all the firms in the sample. For 38 of these firms, there was published indication or proof of involvement in issuing FFS. The classification of a financial statement as false was based on the following parameters: inclusion in the auditors’ report of serious doubts as to the accuracy of the accounts, observations by the tax authorities regarding serious taxation intransigencies which significantly altered the company’s annual balance sheet and income statement, the application of Greek legislation regarding negative net worth, the inclusion of the company in the Athens Stock Exchange categories of “under observation and “negotiation suspended” for reasons associated with the falsification of the company’s financial data and, the existence of court proceedings pending with respect to FFS or serious taxation contraventions. The 38 FFS firms were matched with 38 non-FFS firms. These firms were characterized as non-FFS based on the absence of any indication or proof concerning the issuing of FFS in the auditors’ reports, in financial and taxation databases and in the Athens Stock Exchange.

Results Obtained: The Bayesian Belief Network model achieved the best performance managing to correctly classify 90.3% of the validation sample in a 10-fold cross validation procedure. The accuracy rates of the Neural Network model and the Decision Tree model were 80% and 73.6% respectively. The Type I error rate was lower for all models.

Conclusion: This study investigates the usefulness of Decision Trees, Neural Networks and Bayesian Belief Networks in the identification of fraudulent financial statements. The input vector is composed of ratios derived from financial statements. The three models are
compared in terms of their performances. The results identify the model with the best accuracy rate and highlight the importance of variables in fraudulent financial statement detection. They also indicate that the investigation of financial information can be of use in the identification of FFS and underline the importance of financial ratios.

2.4.8 Kotsiantis et al [Kotsiantis06]

**Purpose:** To explore the effectiveness of machine learning techniques in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS.

**Data Mining Techniques used:** Decision Tree, Artificial Neural Network, Bayesian Network, K - Nearest Neighbour, Support Vector Machines

**Nature of Data Mining Techniques used:** Predictive

**Data Used:** Data Set contained data from 164 Greek listed on the Athens Stock Exchange (ASE) manufacturing firms (no financial companies were included). Auditors checked all the firms in the sample.

For 41 of these firms, there was published indication or proof of involvement in issuing FFS. The classification of a financial statement as false was based on the following parameters: inclusion in the auditors’ report of serious doubts as to the accuracy of the accounts, observations by the tax authorities regarding serious taxation intransigencies which significantly altered the company’s annual balance sheet and income statement, the application of Greek legislation regarding negative net worth, the inclusion of the company in the Athens Stock Exchange categories of “under observation and “negotiation suspended” for reasons associated with the falsification of the company’s financial data and, the existence of court proceedings pending with respect to FFS or serious taxation contraventions. The 41 FFS firms were matched with 123 non-FFS firms. All the variables used in the sample were extracted from formal financial statements, such as balance sheets and income statements.

**Results Obtained:** The K2 algorithm correctly classifies 74.1% of the total sample, 51.2% of the fraud cases and 82.1% of the non-fraud cases. The RBF algorithm manages to correctly classify 73.4% of the total validation sample, 36.6% of the fraud cases and 86.3% of the non-fraud cases. Moreover, C4.5 algorithm succeeds in correctly classifying 85.2% of the fraud cases, 93.3% of the non-fraud cases and 91.2% of the total validation sets.
Furthermore, 3NN algorithm succeeds in correctly classifying 56.1% of the fraud cases, 88.0% of the non-fraud cases and 79.7% of the total validation sets. SMO algorithm correctly classifies 78.66% of the total sample, 48.8% of the fraud cases and 88.6% of the non-fraud cases. Ripper algorithm succeeds in correctly classifying 65.7% of the fraud cases, 94.1% of the non-fraud cases and 86.8% of the total validation set. What is more, logistic regression algorithm manages to correctly classify 75.3% of the total validation sample, 36.6% of the fraud cases and 88.9% of the non-fraud cases.

**Conclusion:** A hybrid decision support system that combines the representative algorithms using a stacking variant methodology have been implemented and achieves better performance than any examined simple and ensemble method. To sum up, this study indicates that the investigation of financial information can be used in the identification of FFS and underline the importance of financial ratios.

**2.4.9 Hoogs et al [Hoogs07]**

**Purpose:** To propose a genetic algorithm approach for detecting financial statement fraud.

**Data Mining Technique used:** Genetic Algorithm

**Nature of Data Mining Technique used:** Predictive

**Data used:** A sample comprising a target class of 51 companies accused by the Securities and Exchange Commission of improperly recognizing revenue and a peer class of 339 companies matched on industry and size (revenue). Variables include 76 comparative metrics, based on specific financial metrics and ratios that capture company performance in the context of historical and industry performance, and nine company characteristics.

**Results Obtained:** Time-based patterns detected by the genetic algorithm accurately classify 63% of the target class companies and 95% of the peer class companies. The genetic algorithm presented in this study misclassified 5% of the non-fraud companies. Classification rates of the approximated model on data sample used in this study were considerably lower than the published classification rates, correctly classifying 43–44% of the alleged fraud companies and misclassifying 22–26% of the non-fraud companies.

**Conclusion:** The exceptional anomaly scores are valuable metrics for characterizing corporate financial behaviour, and that pattern considering the interactions of exceptional anomaly scores over time are effective in detecting potentially fraudulent behaviour.
Authors further concluded by saying that genetic algorithms are a successful technique for detecting discriminatory patterns in challenging domains characterized by high dimensionality and pervasive missing values. The patterns generated by the genetic algorithm are easily translated to domain appropriate language and, therefore, easily understood by external stakeholders. Furthermore, the patterns are capable of identifying potentially fraudulent behaviour despite occasional missing values, and provide low false-positive rates, making them practical for use by external stakeholders.

2.4.10 Belinna et al [Belinna08]

**Purpose:** The objective of this research was to introduce one statistical technique—Classification and Regression Tree (CART), to identify and predict the impacts of FFS.

**Data Mining Technique Used:** Classification and Regression Trees, Logit Regression

**Nature of Data Mining Technique used:** Predictive

**Data used:** The data set for the empirical experiments consists of 148 financial reports. Financial data for the FFS cases are extracted from CSMAR annual report database, and finally form the FFS group with 24 false financial reports. The non-FFS group is formed by searching other available financial reports matched against the industry and the year to the FFS examples, we eliminate those reports with incomplete data and 124 financial reports are included in the non-FFS group.

**Results Obtained:** The performance of two CART experiments and Logit regression were compared. For the original dataset, type I error of CART was 29.17% and that of Logit regression was 37.25%, type II error of CART was 1.61% and that of Logit regression was 4.03%, it is not difficult to observe that CART achieves better accuracy in predicting both the fraud case and non-fraud case with the original dataset; for the benchmarked dataset, type I error of CART was 12.5% and that of Logit regression was 20.83%, type II error of CART was 2.42% and that of Logit regression was 0, although Logit regression achieved better accuracy in predicting the non-fraud case, the error for predicting fraud case was much higher and the overall performance of CART was better (7.46% v. s. 10.42%). Auditors and financial analysts are typically concerned about any possibility of FFS, therefore, an approach with lower type I error rate will better serve this purpose. Overall, CART outperforms Logit regression in both fraud identification accuracy and overall accuracy.
Conclusion: The performance of the proposed approach with Logit regression states that CART outperforms Logit regression, and it is able to produce more accurate classification on the fraud cases, and the prediction accuracy is also higher for CART. Therefore, CART can be considered as an effective approach in FFS identification and prediction, and FFS detection systems can be built based upon such model. However, financial manipulation tricks varies significantly in the sample FFS cases, with a relatively small sample size, the pattern or characteristics cannot be well concluded yet.

2.4.11 Cecchini et al [Cecchini10]

Purpose: To propose a methodology to aid in detecting fraudulent financial reporting by utilizing only basic and publicly available financial data. This study combines aspects of the fraud assessment research in accounting with computational methods and theory used in machine learning / data mining.

Data Mining Techniques used: Support Vector Machine

Nature of data mining technique used: Predictive

Data used: Data sample consist of data of 205 fraudulent companies. The data was gathered by using accounting and auditing enforcement releases (AAERs). The fraud sample was matched with 6,427 nonfraudulent companies. The training sample includes data from 1991 to 2000, and test data set includes data from 2001 to 2003. The training sample, after pre-processing, includes 107 fraud company-years and 2,205 non-fraud company-years. The holdout sample includes 25 fraud company-years and 982 non-fraud company-years.

Results Obtained: Support vector machines using the financial kernel correctly labelled 80% of the fraudulent cases and 90.6% of the non-fraudulent cases on a holdout set. Furthermore, leading fraud research studies were replicated by using the data and finding that the method used in this research has the highest accuracy on fraudulent cases and competitive accuracy on non-fraudulent cases. The results validate the financial kernel together with support vector machines as a useful method for discriminating between fraudulent and non-fraudulent companies using only publicly available quantitative financial attributes. The results also show that the methodology has predictive value because, using only historical data, it was able to distinguish fraudulent from non-fraudulent companies in subsequent years.
Conclusion: This study developed a methodology for detecting management fraud. A domain-specific kernel, the financial kernel, was created that implicitly maps relevant financial attributes to ratios and year-over-year changes of the ratios.

2.4.12 Perols Johan L. [Perols11]

Purpose: To compare the performance of six popular statistical and machine learning models in detecting financial statement fraud under different assumptions of misclassification costs and ratios of fraud firms to non-fraud firms.

Data Mining Techniques used: Logistic Regression, Support Vector Machine, Artificial Neural Network, Decision Trees, Stacking, and Bagging

Nature of Data Mining Techniques used: Predictive

Data used: The fraudulent observations were located based on firms investigated by the SEC for financial statement fraud and reported in Accounting and Auditing Enforcement Releases (AAER) from the fourth quarter of 1998 through the fourth quarter of 2005. A total of 745 potential observations were obtained from this initial search. The data set was then reduced to 272 fraud firms by eliminating: duplicates; financial companies; firms without the first fraud year specified in the SEC release; non-annual financial statement fraud; foreign corporations; releases related to auditors; not-for-profit organizations; and fraud related to registration statements.

Results Obtained: The results show, somewhat surprisingly, that logistic regression and support vector machines perform well relative to an artificial neural network, bagging, C4.5, and stacking. The results also reveal some diversity in predictors used across the classification algorithms. Out of 42 predictors examined, only six are consistently selected and used by different classification algorithms: auditor turnover, total discretionary accruals, Big 4 auditor, accounts receivable, meeting or beating analyst forecasts, and unexpected employee productivity.

Conclusion: Logistic regression, a relatively well-known and established classifier, and SVM outperform or perform as well as a relatively comprehensive set of data mining algorithms. This result is somewhat surprising considering that prior fraud research typically found ANN to either outperform or perform on par with logistic regression. However, this study differs from prior fraud studies in that it evaluates the classifiers using a highly imbalanced dataset, i.e., the minority class has a low prior probability, where the prior minority class probability
is manipulated in both the training and the evaluation data. It also differs from most prior fraud research by examining the performance using optimal classification threshold levels for the different classifiers given a specific evaluation manipulation. Finally, this study differs from prior fraud research that compares classification algorithms by not only including a relatively complete set of attributes, but also using a Wrapper method to select attributes for each classifier. Thus, while the result that logistic regression and SVM outperform or perform as well as the other classifier is somewhat surprising it does not necessarily contradict these prior findings. Rather, the results show that when taking these additional factors into account logistic regression and SVM perform well in the fraud domain. These findings extend financial statement fraud research and can be used by practitioners and regulators to improve fraud risk models.

2.4.13 Ravisankar et al [Ravishankar11]

Purpose: The main objective of this research was to predict the occurrence of financial statement fraud in companies as accurately as possible using intelligent techniques.

Data Mining Techniques Used: Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Genetic Programming (GP), Group Method of Data Handling (GMDH), Logistic Regression (LR), and Probabilistic Neural Network (PNN)

Nature of Data mining Techniques used: Predictive

Data used: The dataset used in this research was obtained from 202 companies that were listed in various Chinese stock exchanges, of which 101 were fraudulent and 101 were non-fraudulent companies. The data also contained 35 financial items for each of these companies.

Results Obtained: PNN with 98.09% accuracy and 98.09% sensitivity outperformed all other classifiers. GP yielded the next best result with 94.14% accuracy and 95.09% sensitivity. PNN is the best classifier among all others in terms of AUC (area under the Receiver Operating Characteristic curve) as well. After using t-statics as a method of feature selection, GP outperformed other classifiers with 92.68% accuracy and 90.55% sensitivity, whereas PNN came close behind with 95.64% accuracy and 91.27% sensitivity. Furthermore, results based on AUC indicated that GP yielded highest accuracy followed by PNN, which yielded marginally less accuracy.
**Conclusion:** It should be noted that while all the techniques have equal cost, the technique that is preferred and recommended is totally dictated by the dataset at hand. Since accuracy is a major concern for financial analysts, one should select that technique which yields less misclassification and consumes less time. This is because the performance of all of these techniques depends on the dataset on which they are used. Having said that, everything else (i.e. accuracies, sensitivity, specificity, etc.) being equal, one should select that technique which is less cumbersome, easy to understand, and easy to implement.

**2.4.14 Other Studies**

[Beasley96] used Logit regression to test the prediction that the inclusion of larger proportions of outside members on the board of directors significantly reduces the likelihood of financial statement fraud with a sample of 150 American firms. They found that non-fraud firms have boards with significantly higher percentages of outside members than fraud firms.

[Hansen96] used a powerful generalized qualitative response model to predict management fraud based on a set of data developed by an international public accounting firm.

[Eining97] conducted an experiment to examine the use of expert systems to enhance the performance of auditors.

[Summers98] constructed a cascaded logit model to investigate the relationship between insider trading and fraud. They found that, in the presence of fraud, insiders reduce their holdings of company stock through high levels of selling activity.

[Busta98] used NN to distinguish between ‘normal’ and ‘manipulated’ financial data. They examined the digit distribution of the numbers in the underlying financial information.

[Deshmukh98] demonstrated the construction of a rule-based fuzzy reasoning system to assess the risk of management fraud and proposed an early warning system by finding out 15 rules related to the probability of management fraud.

[JuszczakOS] apply many different classification techniques in a supervised two-class setting and a semi-supervised one-class setting in order to compare the performances of these techniques and settings.

An innovative fraud detection mechanism is developed by [HuangOS] on the basis of Zipf’s Law. This technique reduces the burden of auditors in reviewing the overwhelming volumes of datasets and assists them in identification of any potential fraud records.

[Cerullo99] explained the nature of fraud and financial statement fraud along with the characteristics of NN and their applications. They illustrated how NN packages could be utilized by various firms to predict the occurrence of fraud.
[Parker00] employed statistical regression analysis to examine if the existence of an independent audit committee mitigates the likelihood of fraud. They found that firms with audit committees, which consist of independent managers who meet at least twice per year, are less likely to be sanctioned for fraudulent or misleading reporting.

[Ngai10] suggested a classification framework for financial fraud detection which consists of two layers. The first comprising the six data mining application classes of classification, clustering, prediction, outlier detection, regression, and visualization supported by second layer which consists of a set of algorithmic approaches to extract the relevant relationships in the data.

[Zhou11] examine the effectiveness and limitations of data mining techniques such as regression, decision trees, and neural network and Bayesian networks. They explore a self-adaptive framework based on a response surface model with domain knowledge to detect financial statement fraud.

[Sharma12] propose a data mining framework for detection of financial fraud. The review of the existing academic literature reveals that the goal of previous research was to identify the financial factors to be used by auditors in assessing the likelihood of financial statement fraud. One main objective was to introduce, apply, and evaluate the use of data mining methods in differentiating organisations between fraud and non-fraud observations. The review further reveals that merely all research is conducted in the field of detection and identification of financial statement fraud. The goal of previous research was just to classify the organisations into fraud or non-fraud by applying statistical tools and predictive data mining techniques.

2.5 SUMMARY

This chapter provides the theoretical framework of data mining techniques and their applicability in prevention and detection of financial statement fraud, which sets the base for further study. Data mining techniques are the most commonly used technique for identification and detection of fraudulent financial reporting.

International Auditing and Assurance Standard Board [IAASB07] states that auditors' main responsibility is to express an opinion about whether financial statements are prepared within an acceptable accounting framework and thus provide assurance that financial statement are free from material misstatement, whether caused by fraud or error. Conventional auditing procedures are not capable enough in preventing and detecting financial statement fraud because in most of the cases auditors are deceived by managers.
Auditors can be assisted in preventing and detecting financial statement fraud by data mining techniques because, these techniques can use past cases of fraud for building models to identify and detect the risk of fraud.

Moreover, data mining methods are helpful in discovering the reasons behind fraudulent financial reporting. Extensive literature survey is done to understand the usage of existing data mining techniques for financial fraud detection in terms of data set used and empirical results. This survey gives insight view of data mining methods and helps in conducting further study.

It is concluded from the survey that all researches have been conducted in the field of identification and detection of financial statement fraud, therefore prevention mechanisms should be focused upon and has become the basis of this study. In addition to this, behavioural characteristics of organisations and descriptive data mining methods are decided to be focused by this study for preventing fraudulent financial reporting.